Neural Process Lett (2016) 44:735–750
DOI 10.1007/s11063-015-9491-9

Learning Informative Features from Restricted Boltzmann Machines

Jakub M. Tomczak[1]

Published online: 1 December 2015

Abstract In recent years deep learning paradigm achieved important empirical success in a number of practical applications such as object recognition, speech recognition and natural language processing. A lot of effort has been put on understanding theoretical aspects of this success, however, still there is no common view on how deep architectures should be trained and thus many open questions remain. One hypothesis focuses on formulating *good criterion* (prior) that may help to learn a set of features capable of disentangling hidden factors. Following this line of thinking, in this paper, we propose to add a penalty (regularization) term to the log-likelihood function that enforces hidden units to maximize entropy and to be pairwise uncorrelated, for given observables. We hypothesize that the proposed framework for learning informative features results in more discriminative data representation that maintains its generative capabilities. In order to verify our hypothesis we apply the regularization term to the Restricted Boltzmann Machine (RBM) and carry out empirical study on three classification problems: character recognition, object recognition, and document classification. The experiments confirm that the proposed approach indeed increases discriminative and generative performance in comparison to RBM trained without any regularization and with the weight-decay, the sparse regularization, the max-norm regularization, *Dropout* and *Dropconnect*.

Keywords Unsupervised learning · Entropy-based regularization · Orthonormality regularization · Restricted Boltzmann machine

1 Introduction

Learning successful classifier or other predictor greatly depends on appropriately prepared data representation. The representation captures latent factors (*features*) explaining varia-

✉ Jakub M. Tomczak
jakub.tomczak@pwr.edu.pl

[1] Wrocław Universtity of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland

tions in data. Obviously, formulating the data representation enforces application of feature engineering, *i.e.*, formulating a set of features using human knowledge and prior information about the considered problem. However, it would be highly desirable to learn features automatically from low-level sensory data, and with possibly no contribution of human hand-tuning [4]. Among many manners of learning representations, deep learning becomes its leading paradigm [3] with applications to various domains, *e.g.*, object recognition [31], speech recognition [16], natural language processing [9], neuroimaging [19].

Explanation of the success of deep learning is still unclear, nonetheless, it is hypothesized that deep architectures help in disentangling factors in data [3]. However, there is no common perspective on how the learning of deep models should be performed. Some insight into the matter has been outlined in [5] and further elaborated in [4] where several general guidelines have been formulated. These guidelines, which were named *general (broad) priors*, point out possible directions in developing new models and learning algorithms. One of such priors is *sparsity of features* [7,12], *i.e.*, only a small fraction of hidden factors are relevant. Sparsity is motivated by biological brain where only 1–4 % of neurons are active at any time [3]. Possible strength of the sparse representation is easier propagation of gradient and better coding of information. However, it is not obvious whether sparsity in shallow and deep models gives the same results or how it influences information coding. Nonetheless, it is believed that sparsity is one the main causes of successful data representations.

Other prior aims at formulating learning objective which enforces robustness of the representation [28]. Recently, an empirical evidence was presented that indeed the deep hierarchy can be helpful in better disentangling the hidden factors [6]. A different question is what type of learning objective results in successful representation. This issue is especially problematic since the representation learning can be performed in the unsupervised manner. For example, in the case of classification the learning objective is rather straightforward, *i.e.*, we want to minimize the number of misclassifications. In the context of unsupervised learning the objective is less obvious, because it is unclear how to formulate the learning objective for providing a representation that contains possibly all information about data and allows simple disentangling underlying factor of variation. In [4] the authors hypothesized that the typical learning objective, *i.e.*, the log-likelihood function, may be expanded by an additional data-dependent regularization term that helps to learn the set of features capable of disentangling factors. Such regularization techniques were proposed and empirically proved to increase quality of the hidden representation, *e.g.*, *Dropout* [35] which is an adaptive (data-dependent) regularizer [36], or its extension *Dropconnect* [37]. In this paper, we will follow this line of thinking and propose new data-dependent regularizers.

The class of deep architectures is wide, however, one of the most popular models that is used as a building block for unsupervised deep representation learning is the Restricted Boltzmann Machine (RBM). The RBM is a Markov random field with a bipartite structure consisting of one layer of observable variables and one layer of hidden units (features). Typically, the RBM is stacked in a hierarchy to build a deep network or it is used as a feature extractor [3]. However, in order to obtain better generalization ability of the RBM, and thus better data representation, learning could be improved by incorporating some *prior* about the world or the representation itself, as mentioned earlier. The prior can be introduced in form of constraints or as a penalty (regularization) term in the learning objective. In the context of deep learning, the widely-used regularization is *weight decay*, *i.e.*, the ℓ_2 norm on parameters [18] that prevents the model from overfitting and helps to stabilize learning process. Other kind of regularization aims at keeping hidden units activation at a constant but small level, which eventually results in sparse representation [24].

Realizing the challenging issue of learning data representation, in this paper we propose a new approach for learning a hidden representation on the example of the RBM. The learning objective for the RBM consists of two main components: the likelihood function and the regularization term that enforces features to maximize entropy and to be pairwise uncorrelated. Particularly, we formulate unconstrained learning problem and further provide information-theoretic regularization, *i.e.*, the sum of entropy of each hidden unit for given observables, and soft orthonormality constraint to obtain uncorrelated features. We call the features trained using these two regularizers *informative features* to highlight the connection of our approach to information theory and binary codes learning [38]. Basing on this formulation of learning informative features, we calculate gradients with respect to weights that can be later used in the stochastic gradient descent learning algorithm. Additionally, we give a close relation of the proposed entropy-based regularization to the variance-based regularization used in learning binary hashing codes [38]. Eventually, we empirically show that the proposed framework for learning informative features results in more discriminative data representation that maintains its generative capabilities measured by the test log-likelihood function.

The remainder of the paper is organized as follows. We first outline the RBM and formulate unconstrained learning problem in Sect. 2. We further present the idea of learning informative features in Sect. 3. In the following subsections we propose the core of our approach, *i.e.*, the information-theoretic regularization (in Sect. 3.1) and the orthonormality regularization and the reconstruction cost (in Sects. 3.2.1 and 3.2.2). In Sect. 4 we relate our proposition to existing approaches. Section 5 gives the experimental results for three datasets, namely, MNIST, CalTech 101 Silhouettes, and 20-newsgroups. At the end, conclusions are drawn in Sect. 7.

2 Restricted Boltzmann Machine

The Restricted Boltzmann Machine (RBM) is a Markov random field that defines the joint distribution over binary visible and hidden units [34]. A distinctive trait of the RBM is its bipartite structure in which visible units (*visibles* or *observables*, for short) and hidden units form two layers and connections within the same layer are prohibited. The relationships among units are specified through the *energy function*:

$$E(\mathbf{x}, \mathbf{h}|\boldsymbol{\theta}) = -\mathbf{x}^\top \mathbf{W}\mathbf{h} - \mathbf{b}^\top \mathbf{x} - \mathbf{c}^\top \mathbf{h}, \tag{1}$$

where $\mathbf{x} \in \{0, 1\}^D$ are the visible units, $\mathbf{h} \in \{0, 1\}^M$ are the hidden units, and $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{b}, \mathbf{c}\}$ is a set of parameters, $\mathbf{W} \in \mathbb{R}^{D \times M}$, $\mathbf{b} \in \mathbb{R}^D$, and $\mathbf{c} \in \mathbb{R}^M$ are, respectively, weights, visible biases, and hidden biases. For the energy function (1) the RBM is defined by the *Gibbs distribution*:

$$p(\mathbf{x}, \mathbf{h}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\big(-E(\mathbf{x}, \mathbf{h}|\boldsymbol{\theta})\big), \tag{2}$$

where $Z(\boldsymbol{\theta})$ is the *partition function*, $Z(\boldsymbol{\theta}) = \sum_{\mathbf{x}} \sum_{\mathbf{h}} \exp\big(-E(\mathbf{x}, \mathbf{h}|\boldsymbol{\theta})\big)$.

Since there are no connections within the same layer, the RBM possesses very useful property that the conditional distribution over the hidden units factorizes given the visible units and the probability of a hidden unit is as follows:[1,2]

$$p(h_m = 1|\mathbf{x}, \boldsymbol{\theta}) = \sigma\big((\mathbf{W}_{\cdot m})^\top \mathbf{x} + c_m\big). \tag{3}$$

[1] Further, in the paper, we use the following notation: for given matrix $\mathbf{A}$, A_{ij} is its element, $\mathbf{A}_{\cdot j}$ denotes its jth column , $\mathbf{A}_{i\cdot}$ denotes its ith row, and for given vector $\mathbf{a}$, a_i is its ith element.

[2] $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the sigmoid function.

Likewise, the conditional distribution over the visible units factorizes given the hidden units and the probability of a visible unit is the following:

$$p(x_d = 1|\mathbf{h}, \boldsymbol{\theta}) = \sigma(\mathbf{W}_{d.}\mathbf{h} + b_d). \tag{4}$$

2.1 Learning

We assume given N training data which are organized in a data matrix $\mathbf{X} = [x_{dn}]_{D\times N}$, where each column is a data point $\mathbf{x}_n$. Training the RBM corresponds to minimizing the negative log-likelihood of N data with respect to parameters $\{\mathbf{W}, \mathbf{b}, \mathbf{c}\}$:

$$\mathcal{L}(\boldsymbol{\theta}) = -\sum_n \log p(\mathbf{x}_n|\boldsymbol{\theta}), \tag{5}$$

where $p(\mathbf{x}_n|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})}\exp\{-FE(\mathbf{x}_n|\boldsymbol{\theta})\}$ and $FE(\mathbf{x}|\boldsymbol{\theta})$ is a *free energy*:

$$FE(\mathbf{x}|\boldsymbol{\theta}) = -\mathbf{b}^\top\mathbf{x} - \sum_j \log\left\{1 + \exp\left\{c_j + (\mathbf{W}_{.j})^\top\mathbf{x}\right\}\right\}. \tag{6}$$

In general, the gradient of (5) cannot be computed analytically because of the partition function. However, it can be efficiently approximated using the contrastive divergence [17] or other inductive principle [25].

The main challenge in automatic feature learning is to design learning algorithms that can discover representations that compactly characterize regularities in data [5]. One possible fashion of obtaining appropriate representation is to formulate a *general* or *broad prior* [4,5] that reduces the space of accessible functions, such that, enforcing *smoothness*, or *sparsity* [4]. In general, the prior can be cast in form of a penalty (or regularization) term $\Omega(\boldsymbol{\theta})$, thus, the learning problem can be stated as an unconstrained optimization problem as follows:

$$\text{minimize (w.r.t. } \boldsymbol{\theta})\mathcal{L}_\Omega(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}) + \lambda\Omega(\boldsymbol{\theta}) \tag{7}$$

where $\lambda > 0$ is a penalty (or regularization) coefficient.

In the context of deep models, an example of such approach is a regularization technique known as *weight decay* which controls the ℓ_2 norm of the parameters and leads to smooth models:[3]

$$\Omega_{\text{wd}}(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{W}\|_F^2. \tag{8}$$

3 Learning Informative Features

Learning data representation is conceptually closely related to learning binary codes where a single bit can be associated with a feature. In the learning binary codes, ideally one should propose a code-book that encodes the reality such that the information provided by each bit is maximized. In other words, the entropy of each bit for given data is maximized. Additionally, one would like to avoid redundancy in bits as much as possible that is equivalent to bits that are pairwise uncorrelated. We will use these two principles for learning data representation of the RBM by introducing constraints enforcing entropy of hidden units to be maximized and probabilities of hidden units to be pairwise uncorrelated. We call features trained in this manner *informative features* to highlight that our motivation is anchored in the theory of

[3] $\|\cdot\|_F$ is the Frobenius norm.

learning binary codes [14,33,38,39]. The learning binary codes problem is formulated as a constrained optimization problem with two hard constraints, namely, balancedness of hidden units (*i.e.*, maximization of entropy of hidden representation) and orthonormality of columns in the matrix of weights (a relaxed version of the pairwise decorrelation of hidden units).

In the context of the RBM we believe that the entropy constraint could improve mixing of Markov chain during training because the Markov chain started from a training example can easier change the corresponding hidden states. Additionally, we hope that the balancedness of hidden units can make it harder for the classifier using the features from the RBM to overfit and thus can improve the discriminative capabilities of the representation. Furthermore, we expect that application of the constraint enforcing decorrelation of the probabilities of hidden units results in features for which better separability of classes can be achieved.

3.1 Information Theoretic Regularization

In general, direct application of the balancedness of hidden units, *i.e.*, the entropy constraint, could be inaccessible. Therefore, we propose to introduce a regularization term that enforces entropy of the hidden units to be maximized:[4]

$$\Omega_{\mathrm{it}}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_n\sum_m \mathcal{H}(h_m|\mathbf{x}_n) \tag{9}$$

$$= \frac{1}{N}\sum_n\sum_m \Big(\sigma_{mn}\log(\sigma_{mn}) + (1-\sigma_{mn})\log(1-\sigma_{mn})\Big), \tag{10}$$

where $\sigma_{mn} \triangleq p(h_m = 1|\mathbf{x}_n, \boldsymbol{\theta})$. We refer $\Omega_{\mathrm{it}}(\boldsymbol{\theta})$ to as the *information theoretic regularization*.

Further in the experiments we apply stochastic gradient descent algorithm for learning parameters. Therefore, we need to calculate derivative w.r.t. to a weight W_{dm} that yields:

$$\frac{\partial}{\partial W_{dm}}\Omega_{\mathrm{it}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_n \sigma_{mn}(1-\sigma_{mn}) \times (\log\sigma_{mn} - \log(1-\sigma_{mn}))x_{dn}$$

$$= \frac{1}{N}\sum_n v_{mn}x_{dn}, \tag{11}$$

where, for brevity, we define

$$v_{mn} \triangleq \sigma_{mn}(1-\sigma_{mn})(\log\sigma_{mn} - \log(1-\sigma_{mn})). \tag{12}$$

The quantity v_{mn} can be seen as a modified variance of the m-th hidden unit for n-th example. Modifying the variance $\sigma_{mn}(1-\sigma_{mn})$ by the non-linear term $\log(\sigma_{mn}) - \log(1-\sigma_{mn})$ in (11) causes penalization of probabilities different than $\frac{1}{2}$. This effect is desirable because we would like to obtain hidden units which maximize entropy. Additionally, notice that the weight is penalized only if the d-th value of n-th example is 1, otherwise no regularization is applied. Such dependence on data could be advantageous since for inactive visible unit, $x_{dn} = 0$, the probability of hidden unit activation relies only on the bias (see 3) and thus the weight should not be updated.

[4] We aim at minimizing the learning objective which is a sum of the negative log-likelihood and the penalty (regularization) term, see (7), while we want to maximize the entropy. Hence, we need to include minus in the regularization term.

For the sake of completeness, since we have calculated derivative w.r.t. to single weight, we give the gradient w.r.t. to the weight matrix:

$$\nabla_{\mathbf{W}} \Omega_{\text{it}}(\boldsymbol{\theta}) = -\frac{1}{N} \mathbf{X}\mathbf{V}^\top. \tag{13}$$

where $\mathbf{V} = [v_{mn}]_{M \times N}$.

As a result we obtained a data-dependent weights update. The weights are updated using both the data matrix and by incorporating the matrix of modified variances of hidden units that are dependent on the parameters.

At the end, we would like to point out a connection between the information-theoretic regularization and the weight decay. The proposed entropy-based regularization aims at finding weights that maximize the entropy *explicitly*. On the other hand, the application of the weight decay enforces weights to be close to zero. As a result, since $\mathbf{W}$ is pulled towards $\mathbf{0}$, the probability of the hidden unit given observables becomes $\frac{1}{2}$.[5] Therefore, one gets the value of the probability that maximizes entropy. Therefore, we notice that the weight decay maximizes the entropy *implicitly*.

3.2 Orthonormality Constraint

The constraint on the probabilities of hidden units being pairwise uncorrelated can be relaxed to the condition on orthonormality of columns in the matrix of weights, *i.e.*, the orthonormality of the projection directions [14,38]:

$$\mathbf{W}^\top \mathbf{W} = \mathbf{I}. \tag{14}$$

However, imposing hard orthonormality constraint could lead to several sever problems. For example, if $\mathbf{W}$ is overcomplete, *i.e.*, $D < M$, then the constraint can no longer be satisfied [22]. Moreover, considering orthogonal directions does not necessarily lead to informative and discriminative features [38]. Therefore, the hard constraint can be introduced as a regularization term to overcome these issues. In the following sections we present two possible manners of introducing the hard orthonormality constraint as the penalty term.

3.2.1 Orthonormality Regularization

One possible fashion of incorporating the orthonormality constraint in the learning problem is to include the *orthonormality regularization* term in the following form [38]:

$$\Omega_{\text{o}}(\boldsymbol{\theta}) = \frac{1}{2} \|\mathbf{W}^\top \mathbf{W} - \mathbf{I}\|_F^2. \tag{15}$$

Such regularization term allows to select which projection directions should be orthonormal by balancing various terms. The gradient of (15) w.r.t. weights is as follows:

$$\nabla_{\mathbf{W}} \Omega_{\text{o}}(\boldsymbol{\theta}) = \mathbf{W}(\mathbf{W}^\top \mathbf{W} - \mathbf{I}). \tag{16}$$

3.2.2 Reconstruction Cost

The biggest problem with the orhonormality constraint is that it fails for overcomplete matrices. This observation served as a starting point for the authors of [22] to formulate a new regularizer for the non-degeneracy control of the weights matrix:

[5] Neglecting the bias term c_m in (3) for simplicity.

$$\Omega_{\mathrm{rc}}(\boldsymbol{\theta}) = \frac{1}{2N}\sum_{n=1}^{N} \|\mathbf{W}^{\top}\mathbf{W}\mathbf{x}_n - \mathbf{x}_n\|_2^2. \tag{17}$$

which is known as the *reconstruction cost*. Notice, that in this regularizer it is assumed the weights are tied, *i.e.*, $\mathbf{W} = \mathbf{W}^{\top}$. In [22] it has been shown that the reconstruction cost is equivalent (for certain conditions) to the orthonormality constraint for ICA, autoencoders and sparse coding. This is another example of the data-dependent regularizer.

The gradient of (17) w.r.t. weights is the following:

$$\nabla_{\mathbf{W}}\Omega_{\mathrm{rc}}(\boldsymbol{\theta}) = \frac{1}{N}\mathbf{X}\mathbf{X}^{\top}\mathbf{W}(\mathbf{W}^{\top}\mathbf{W} - \mathbf{I}). \tag{18}$$

It turns out that the gradient of the reconstruction cost is a modified version of the result obtained for the orthonormality regularization where the gradient (16) is weighted by the data covariance matrix.

3.3 Remarks

The gradients of the information-theoretic regularization and the reconstruction cost given by (13) and (18), respectively, require summing over the whole training set that may be troublesome. Therefore, to alleviate this issue, mini-batches will be used in the experiments.

In our considerations we focused on calculating gradients w.r.t. to weights only. It is a common practice not to regularize biases because they are less likely to cause overfitting and sometimes they even need to be large [18].[6]

4 Related Works

The idea of exploiting entropy as a regularization term is well known in machine learning. However, in the considered case we introduce entropy-based regularization in order to obtain specific property of the hidden representation while typically the entropy-based regularization is used in other context, *e.g.*., in the semi-supervised learning [10,15,27] or supervised regularization of deep models [13].

The idea of our approach is closely related to the one utilized in learning binary hashing codes [14,38]. However, there are two major differences. First, sometimes the entropy of the binary code serves as the objective [14] while in our case it is a constraint (similarly as it is done in [38,39]). Second, in learning binary hashing codes each bit is described by a signum function and thus direct calculation of derivatives w.r.t. to weights is prohibited and some kind of relaxation of the regularization term is needed. The authors of [38] proposed a lower bound of the sum of variances (since maximum of bit entropy and variance coincide and both quantities are convex). However, we deal with sigmoids and thus the derivative of the regularizer can be calculated exactly (see Eq. 11). Nonetheless, it could be tempting to follow the same line of thinking as in [38] to re-express the balancedness constraint in terms of variance instead of entropy that leads to the following regularizer:

[6] In the preliminary experiments we tested regularizing biases too and indeed such approach resulted in worst results.

$$\Omega_{\mathrm{v}}(\boldsymbol{\theta}) = -\frac{1}{2N}\sum_n\sum_m \mathrm{Var}(h_m|\mathbf{x}_n) \tag{19}$$

$$= -\frac{1}{2N}\sum_n\sum_m \sigma_{mn}(1-\sigma_{mn}). \tag{20}$$

Further, calculating the derivative w.r.t. to W_{dm} gives:

$$\frac{\partial}{\partial W_{dm}}\Omega_{\mathrm{v}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_n \sigma_{mn}(1-\sigma_{mn})\times (2\sigma_{mn}-1)x_{dn}. \tag{21}$$

Taking a closer look at (11) and (21) one immediately notices that the difference in the derivatives is how the variances are modified. Applying the entropy-based regularization results in non-linear modification while the variance-based regularization outcomes linear modification.

The hard orthonormality constraint is usually used in the context of learning binary codes [14,38,39]. However, as pointed out by the authors of [22], this constraint is also utilized in the problem statement of the Independent Component Analysis (ICA). Moreover, in [22] the reconstruction cost is proposed as a soft orthonormality constraint that results in a new formulation of the ICA. However, in our work we apply the reconstruction cost to obtain specific property of hidden representation of a deep model (RBM) instead of learning data representation directly as in the ICA.

5 Experiments

5.1 Datasets

We present empirical evaluation of the proposed approach on three classification problems: character recognition problem using MNIST[7], object recognition problem using CalTech 101 Silhouettes Data Set[8], and the document classification using 20-newsgroup dataset[9].

The MNIST dataset [23] contains images of 28×28 pixels of ten hand-written digits (from 0 to 9). The dataset is divided into 50,000 training images, a validation set of 10,000 examples and 10,000 images are used for testing.

The CalTech 101 Silhouettes Data Set [25], CalTech for short, consists of 28×28-size images representing black silhouettes of 101 objects (classes) on white background. The training set contains 4100 images with at least 20, and at most 100 examples from each class. The remaining examples are split into a validation set and test set of size 2264 and 2307, respectively.

The 20-newsgroups dataset [20], 20Newsgroups for brevity, contains 8500 training, 1245 validation, and 6497 test documents. We used 100 most frequent words describing a document as the binary inputs. The problem is to classify a document to one of four newsgroup meta-topics (classes).

[7] Data taken from: http://yann.lecun.com/exdb/mnist/.

[8] Data taken from: https://people.cs.umass.edu/~marlin/data.shtml.

[9] In the experiments we used the small version of the original dataset: http://www.cs.nyu.edu/~roweis/data.html.

5.2 Evaluation methodology

In order to verify the effect of applying the entropy and orthonormality constraints together and alone as a penalty term, in the experiments we compare the following learning schema:

- learning of RBM without any regularizer (denoted by RBM);
- learning of RBM with the weight decay, Ω_{wd} (RBM+wd);
- learning of RBM with the information-theoretic regularization, Ω_{it} (RBM+itr);
- learning of RBM with the orthonormality regularization, Ω_{o} (RBM+ortho);
- learning of RBM with the reconstruction cost, Ω_{rc} (RBM+rc);
- learning of RBM with the weight decay and the orthonormality regularization, $\Omega_{\text{wd}} + \Omega_{\text{o}}$ (RBM+wd+ortho);
- learning of RBM with the weight decay and the reconstruction cost, $\Omega_{\text{wd}} + \Omega_{\text{rc}}$ (RBM+wd+rc);
- learning of RBM with the information-theoretic regularization and the orthonormality regularization, $\Omega_{\text{it}} + \Omega_{\text{o}}$ (RBM+itr+ortho);
- learning of RBM with the information-theoretic regularization and the reconstruction cost, $\Omega_{\text{it}} + \Omega_{\text{rc}}$ (RBM+itr+rc).

For any regularization term, *i.e.*, single penalty term or a sum of two regularizers, we used common regularization coefficient. Moreover, we compared the proposed regularization techniques with the following regularizers:

- *sparsity regularization* a regularizer that aims at forcing the RBM to activate only a fixed number of hidden units [26];
- *max-norm regularization* a regularizer that reduces weight values if the norm of a weight vector for given hidden unit exceeds a given value (see Appendix A.3 in [35] for details);
- *Dropout* randomly dropping hidden units (along with their connections) with probability 0.5 during training [35];
- *Dropconnect* randomly dropping connections between input and hidden units with probability 0.5 during training [37];

Additionally, we evaluated the classification restricted Boltzmann machine (ClassRBM) trained in a generative fashion [21] in the classification task.

There are two possible approaches to verify the discriminative capability of the hidden representation. The trained model like the RBM can be further used as an initialization of a feedforward neural network. This is known as a *pre-training* and was shown to act as a very specific data-dependent regularizer used only once during learning [11]. The other approach assumes that the hidden representation is used for further training of a (linear) classifier [8]. If the classifier obtains high discriminative performance, then we can say that the hidden features are discriminative. In this paper, we decided to use the hidden representation directly for further learning of a classifier because in the former approach we may be confused with the data-dependent regularization effect.

The performance of the considered learning schema was measured using three evaluation metrics: test classification error, test log-likelihood, and average number of active hidden units. The test classification error determines the discriminative performance of the trained hidden representation (features) given as an input to the linear classifier. The test log-likelihood of the RBM was calculated with the approximation of the normalizing constant using Annealed importance sampling technique [30]. This metric is used to evaluate the generative capability of the trained RBM. The average number of active hidden units is calculated in order to verify whether application of the considered regularizers causes increase

Table 1 Test classification error for different learning schema and three datasets considered in the experiments

Model	Classification error (%)		
	MNIST	CalTech	20 Newsgroups
RBM	3.25 ± 0.01	35.81 ± 0.20	19.59 ± 0.05
RBM+wd	2.90 ± 0.01	34.61 ± 0.15	19.27 ± 0.04
RBM+itr	3.13 ± 0.02	34.29 ± 0.16	19.32 ± 0.05
RBM+ortho	2.9 ± 0.01	36.23 ± 0.19	18.96 ± 0.03
RBM+rc	2.89 ± 0.01	34.07 ± 0.14	18.77 ± 0.03
RBM+wd+ortho	2.76 ± 0.02	33.68 ± 0.16	18.92 ± 0.04
RBM+wd+rc	2.65 ± 0.01	32.99 ± 0.18	18.92 ± 0.04
RBM+itr+ortho	2.73 ± 0.01	33.85 ± 0.17	19.10 ± 0.05
RBM+itr+rc	$\mathbf{2.46} \pm 0.01$	$\mathbf{32.60} \pm 0.16$	$\mathbf{18.73} \pm 0.03$
RBM+sparse	2.56 ± 0.01	34.9 ± 0.16	19.64 ± 0.05
RBM+max-norm	3.65 ± 0.03	41.78 ± 0.22	19.18 ± 0.04
RBM+Dropout	3.79 ± 0.04	33.29 ± 0.21	19.56 ± 0.06
RBM+Dropconnect	5.61 ± 0.06	37.28 ± 0.25	18.99 ± 0.07
ClassRBM	5.32 ± 0.01	40.63 ± 0.18	18.98 ± 0.02

The best results are in bold

of activity of hidden units. Since we aim at obtaining uncorrelated features such increase of activity is rather expected.

All reported results are averaged over three repetitions of the experiment.

5.3 Training protocol

We perform learning RBMs using the contrastive divergence with mini-batches of 10 examples. We used 500 hidden units for all datasets. The learning rate was set using the model selection for the following values: $\{10^{-1}, 10^{-2}, 10^{-3}\}$. Additionally, we set the regularization coefficient in (7) according to the model selection for $\lambda \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. The number of iterations over the training set was determined using early stopping according to the validation set reconstruction cross-entropy error, with a look ahead of 5 epochs. In all experiments for each dataset the same initialization of the parameters was applied.

Further, for classification the linear classifier, namely, the logistic regression[10], was fed up with the probabilities of hidden units, $p(h_j = 1|\mathbf{x})$, as inputs.

6 Results and discussion

The results of the experiments are presented in Tables 1, 2, and 3 for the test classification error, test log-likelihood, and average number of active hidden units, respectively. In Fig. 1 the exemplary weights of the RBM trained without any regularization term and with the entropy-based regularizer and the reconstruction cost are presented. In Fig. 2 the number of epochs during learning process for different learning schema is depicted. For clarity of visual comparison, we have omitted results of the orthonormality regularization (RBM+ortho),

[10] The ℓ_2 regularization on parameters was applied with the regularization coefficient equal $\{10^0, 10^{-1}, 10^{-2}\}$.

Table 2 Test log-likelihood for different learning schema and datasets considered in the experiments

Model	Test log-likelihood		
	MNIST	CalTech	20Newsgroups
RBM	-131.19 ± 0.13	-216.52 ± 0.21	$\mathbf{-13.79} \pm 0.12$
RBM+wd	-130.46 ± 0.12	-173.49 ± 0.19	-13.94 ± 0.11
RBM+itr	-131.59 ± 0.19	-169.15 ± 0.20	-13.89 ± 0.12
RBM+ortho	-126.54 ± 0.14	-144.43 ± 0.18	-14.87 ± 0.15
RBM+rc	-153.03 ± 0.16	-132.54 ± 0.21	-14.92 ± 0.16
RBM+wd+ortho	-126.08 ± 0.12	-172.67 ± 0.20	-14.82 ± 0.14
RBM+wd+rc	-150.27 ± 0.15	-156.55 ± 0.19	-14.31 ± 0.13
RBM+itr+ortho	-125.85 ± 0.12	-199.77 ± 0.19	-14.95 ± 0.15
RBM+itr+rc	$\mathbf{-118.81} \pm 0.13$	$\mathbf{-129.75} \pm 0.19$	-14.91 ± 0.13
RBM+sparse	-127.93 ± 0.11	-177.46 ± 0.20	-13.89 ± 0.14
RBM+max-norm	-158.17 ± 0.13	-238.78 ± 0.25	-14.04 ± 0.13
RBM+Dropout	-363.36 ± 1.25	-190.91 ± 0.22	-15.95 ± 0.13
RBM+Dropconnect	-213.74 ± 0.83	-264.87 ± 0.24	-14.85 ± 0.12

The best results are in bold

Table 3 Average number of active hidden units for different learning schema and datasets considered in the experiments

Model	Avg. number of active hidden units		
	MNIST	CalTech	20 Newsgroups
RBM	37	91	1
RBM+wd	39	89	1
RBM+itr	37	90	1
RBM+ortho	48	90	1
RBM+rc	50	130	1
RBM+wd+ortho	77	120	2
RBM+wd+rc	52	130	2
RBM+itr+ortho	78	128	2
RBM+itr+rc	65	128	2
RBM+sparse	36	88	1
RBM+max-norm	44	62	1
RBM+Dropout	61	153	2
RBM+Dropconnect	19	97	1

In all cases the standard deviation was equal around 1

the reconstruction cost (RBM+rc), the weight decay with the orthonormality regularization (RBM+wd+ortho), and the weight decay with the reconstruction cost (RBM+wd+rc), the sparsity and the max-norm regularization, *Dropout* and *Dropconnect* from Fig. 2. It is worth noting here that until now *Dropout* and *Dropconnect* were mainly analyzed in the context of the feedforward neural nets and the convolutional nets. For the RBM only some qualitative conclusions were given for the *Dropout* that it leads to sparse representation [35], however, due to our knowledge no quantitative results for the RBM are reported.

We notice that application of the information theoretic regularization alone increases the discriminative performance of the data representation comparing to the RBM trained without any regularization. However, better improvement can be seen in the case of the orthonormality regularization and the reconstruction cost. Though, the biggest decrease in classification

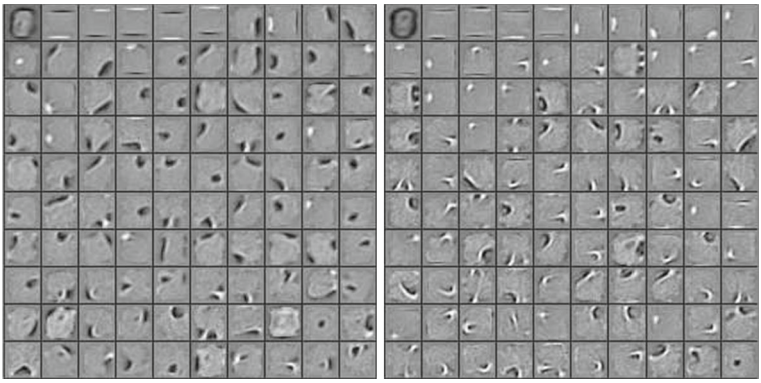

Fig. 1 Exemplary learned weights and visible biases on the MNIST benchmark dataset for the RBM trained without any regularization term (*left*) and the one trained with the best performing regularizer, namely, RBM+itr+rc (*right*). The *top left* cell in each figure is the visible bias vector

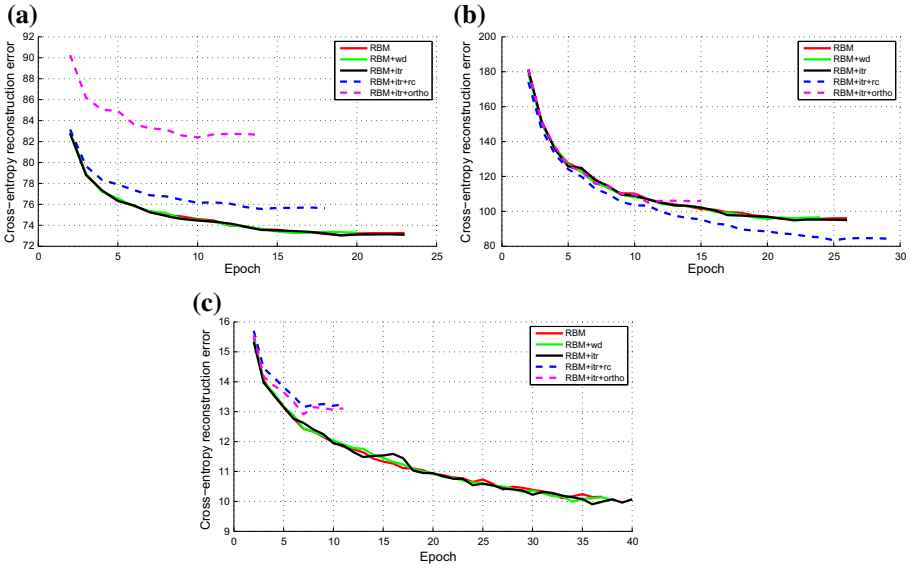

Fig. 2 Number of epoch during learning process for the considered learning schema. For easy of visual comparison, we omitted results for RBM+ortho, RBM+rc, RBM+wd+ortho, RBM+wd+rc, the sparse regularization, the max-norm regularization, *Dropout* and *Dropconnect*

error was achieved by the combination of both the information theoretic regularization and the reconstruction cost. This result reveals that indeed both constraints on the data representation are significant in learning informative features. Slightly worst but still very good performance was obtained by the weight decay with the reconstruction cost. This result is not surprising because the weight decay maximizes entropy implicitly, as pointed out in Sect. 3.1. Additionally, we observe that application of the reconstruction cost may be superior to the orthonormality regularization (see eight and ninth rows of Table 1). Quite good results were obtained by the sparse regularization (second best on MNIST and slightly better than the RBM on CalTech), however, still worst than the RBM+itr+rc. The *Dropout* worked similarly to the ordinary RBM with almost no positive effect (the only except was CalTech dataset).

The max-norm regularization and the *Dropconnect* gave better results than the RBM only on 20Newsgroups but they completely failed on MNIST and CalTech. Surprisingly, ClassRBM performed very badly in comparison to the logistic regression trained with the features extracted from the RBM. However, this effect has been already noticed, *e.g.*, ClassRBM requires 6000 of hidden units to reach the level of error equal 3.39 on MNIST [21].

In the context of generative evaluation, application of any of the considered regularizers alone did not result in any improvement. However, the combination of the information theoretic regularization and the reconstruction cost, similarly to the discriminative performance, led to the best outcome measured in terms of the test log-probabilities (on MNIST and CalTech datasets, see Table 2). This effect was not encountered on the 20Newsgroups dataset where the RBM without any regularization performed the best but it can be explained in the following manner. It may happen that the test log-likelihoods become smaller because smoothing a contribution from each hidden unit could potentially increase the peaks around training samples which further results in lower probability being assigned to nearby test examples. However, the difference between the RBM and the RBM+itr+rc on the 20Newsgroups is slight and hence it is difficult to draw any decisive conclusion. Interestingly, the sparse regularization helps to obtain better generative performance than the ordinary RBM but is is still worst than the RBM+itr+rc. The max-norm regularization fails completely, similarly to the *Dropout* and the *Dropconnect*. This may be a reason why these regularizers are useful in the feedforward neural networks but are rarely used in the context of the RBM.

In order to get a better understanding of how the considered regularizers affect the data representation, we also examined the average number of active hidden units. As stated earlier, we expected that the application of any of the regularizers increases the activity of hidden units and indeed the experiments confirmed our presumptions partially (see Table 3). It seems that application of the information theoretic regularization does not influence the activation of the hidden units. In the case of the orthonormality regularization it is hard to conclude (slight increase on MNIST only, see Table 3). However, for RBM+rc, RBM+itr+ortho, RBM+itr+rc, RBM+wd+ortho and RBM+wd+rc the effect of increase of hidden units activation is apparent. The reason why the value of active hidden units is increased follows directly from the applied regularizers. The proposed regularization terms imitate how the binary codes are trained, *i.e.*, each bit (a feature) tries to maximize information in data while a code (the whole representation) avoids redundancy in bits. As a consequence, the more two observables differ (in the sense of Hamming distance), the more distinct features (hidden units) switch on. Therefore, we notice increased value of active hidden units for the proposed regularizers because possibly the trained representation activates a number of *common* features but also a number of *specific* hidden units for given data. We leave inspecting this issue for future research.

We also analyzed a potential influence of the considered regularizers on the number of epochs in the learning process. It turned out that the application of the soft orthonormality constraint reduced the number of learning iterations (the only exception was RBM+itr+rc on CalTech, see Fig. 2b). This effect is especially evident on 20Newsgroups dataset where the number of epochs was dramatically smaller for RBM+itr+rc and RBM+itr+ortho in comparison to other learning schema (see Fig. 2c). The possible explanation of this result may arise from the fact that the informative features allow better mixing of Gibbs sampler in the contrastive divergence procedure. Since entropy and orthonormality constraints enforce the probabilities of hidden units to be balanced and uncorrelated, different set of features should be activated for given different visibles. Therefore, the blocked Gibbs sampling procedure can easier mix between modes and thus the learning process needs less iterations. However, this claim needs further investigation.

Eventually, as a qualitative assessment of the trained RBM models, we show trained weights and visible biases for the ordinary RBM and the one trained with the information-theoretic regularization and the reconstruction cost on the MNIST benchmark dataset. The weights with highest value of ℓ_2 norm are shown in Fig. 1 where the top left cell in each figure is the visible bias vector. We can easily verify that the application of the proposed regularization leads to different filters, *i.e.*, spot filters, and strokes filters.

7 Conclusion

In this paper, we proposed to add the regularization term to the log-likelihood function of the RBM that enforces hidden units to maximize entropy and to be pairwise uncorrelated, for given visibles. Specifically, we formulated unconstrained penalized learning problem and provided the penalty as a sum of the information-theoretic regularization, *i.e.*, the sum of entropy of each hidden unit for given observables, and the soft orthonormality constraint in terms of the orthonormality regularization or the reconstruction cost to obtain uncorrelated features. We evaluated our approach on the example of the RBM, a well-studied building block of deep models. In the experiments we provided empirical evidence that the proposed regularization led to learning better discriminative data representation than the one obtained without any regularization and with other considered regularizers. Moreover, we noticed that the RBM trained with the combination of the information-theoretic regularization and the reconstruction cost achieved the best discriminative and generative performance among the considered learning schema. Additionally, the RBM with the application of the soft entropy and orthonormality constraints reduced the number of iterations of the learning process.

For future research, we would like to investigate the use of the proposed regularization term in pre-training of deep networks. The results for single RBM are very promising, hence we believe that the effect of learning informative features can be more evident in deeper architectures. Especially, since we aim at learning uncorrelated hidden units, a deep model can easier disentangle factors at consecutive levels. Moreover, the application of orthonormality regularization or the reconstruction cost increases hidden units activation. This results is in contrary to the common opinion that deep models should be sparse [12]. We find that combining sparsity with learning informative features is potentially an interesting research direction. Moreover, in the experiment we found that application of the soft entropy and orthonormality constraints lead to reduction of the number of learning iterations. Our hypothesis is that the informative features allow better mixing of the Gibbs sampler in the contrastive divergence procedure. This statement requires more thorough investigation. In the experiments we assumed a fixed number of hidden units. However, there is plenty of works on how to choose proper structure of a model using the model selection with different complexity penalty terms [2], naming only a few, Akaike's information theoretic criterion [1,32] or Rissanen's minimum description length [29], Last but not least, in this work we considered a probabilistic deep model. However, there is a vast of deterministic deep models based on auto-encoders. Since exact application of the proposed approach is impossible in the context of auto-encoders, it is challenging to suggest corresponding fashion of learning informative features.

Acknowledgments The author is grateful to Adam Gonczarek and Maciej Zięba for fruitful discussions and anonymous reviews for helpful comments. The research conducted by the author has been partially co-financed by the Ministry of Science and Higher Education, Republic of Poland, Grant No. B40020/I32.

References

1. Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723
2. Arribas JI, Cid-Sueiro J (2005) A model selection algorithm for a posteriori probability estimation with neural networks. Neural Netw IEEE Trans 16(4):799–809
3. Bengio Y (2009) Learning deep architectures for AI. Found Trends Mach Learn 2(1):1–127
4. Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828
5. Bengio Y, LeCun Y (2007) Scaling learning algorithms towards AI. Large-scale Kernel Mach 34:1–41
6. Bengio Y, Mesnil G, Dauphin Y, Rifai S (2013) Better mixing via deep representations. In: ICML, pp 552–560
7. Cho K (2013) Simple sparsification improves sparse denoising autoencoders in denoising highly corrupted images. In: Proceedings of the 30th international conference on machine learning (ICML-13), pp 432–440
8. Coates A, Ng AY, Lee H (2011) An analysis of single-layer networks in unsupervised feature learning. In: International conference on artificial intelligence and statistics, pp 215–223
9. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P (2011) Natural language processing (almost) from scratch. J Mach Learn Res 12:2493–2537
10. Corduneanu A, Jaakkola T (2002) On information regularization. In: Proceedings of the Nineteenth conference on uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc, pp 151–158
11. Erhan D, Bengio Y, Courville A, Manzagol PA, Vincent P, Bengio S (2010) Why does unsupervised pre-training help deep learning? J Mach Learn Res 11:625–660
12. Glorot X, Bordes A, Bengio Y (2011) Deep sparse rectifier networks. In: Proceedings of the 14th international conference on artificial intelligence and statistics, vol. 15, pp 315–323
13. Goh H, Thome N, Cord M, Lim JH (2013) Top-down regularization of deep belief networks. In: NIPS, pp 1878–1886
14. Gong Y, Lazebnik S, Gordo A, Perronnin F (2013) Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval. Pattern Anal Mach Intell IEEE Trans 35(12):2916–2929
15. Grandvalet Y, Bengio Y (2005) Semi-supervised learning by entropy minimization. In: Advances in neural information processing systems, pp 529–536
16. Graves A, Mohamed A, Hinton G (2013) Speech recognition with deep recurrent neural networks. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 6645–6649
17. Hinton GE (2002) Training products of experts by minimizing contrastive divergence. Neural Comput 14(8):1771–1800
18. Hinton GE (2012) A practical guide to training restricted Boltzmann machines. In: Neural networks: tricks of the trade, Springer, pp 599–619
19. Hjelm RD, Calhoun VD, Salakhutdinov R, Allen EA, Adali T, Plis SM (2014) Restricted Boltzmann machines for neuroimaging: an application in identifying intrinsic networks. NeuroImage 96:245–260
20. Lang K (1995) Newsweeder: learning to filter netnews. In: ICML, pp 331–339
21. Larochelle H, Bengio Y (2008) Classification using discriminative restricted boltzmann machines. In: Proceedings of the 25th international conference on machine learning, pp 536–543
22. Le QV, Karpenko A, Ngiam J, Ng AY (2011) ICA with reconstruction cost for efficient overcomplete feature learning. In: NIPS, pp 1017–1025
23. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
24. Lee H, Ekanadham C, Ng AY (2008) Sparse deep belief net model for visual area V2. In: NIPS, pp 873–880
25. Marlin BM, Swersky K, Chen B, Freitas ND (2010) Inductive principles for restricted boltzmann machine learning. In: International conference on artificial intelligence and statistics, pp 509–516
26. Nair V, Hinton GE (2009) 3d object recognition with deep belief nets. In: Advances in neural information processing systems, pp 1339–1347

27. Niu G, Dai B, Yamada M, Sugiyama M (2012) Information-theoretic semisupervised metric learning via entropy regularization. In: International conference on artificial intelligence and statistics, pp 509–516
28. Rifai S, Vincent P, Muller X, Glorot X, Bengio Y (2011) Contractive auto-encoders: explicit invariance during feature extraction. In: Proceedings of the 28th international conference on machine learning (ICML-11), pp 833–840
29. Rissanen J (1983) A universal prior for integers and estimation by minimum description length. Ann Stat, 416–431
30. Salakhutdinov R, Murray I (2008) On the quantitative analysis of deep belief networks. In: Proceedings of the international conference on machine learning, vol. 25
31. Salakhutdinov R, Tenenbaum JB, Torralba A (2013) Learning with hierarchical-deep models. IEEE Trans Pattern Anal Mach Intell 35(8):1958–1971
32. Seghouane AK, Amari SI (2007) The AIC criterion and symmetrizing the Kullback-Leibler divergence. IEEE Trans Neural Netw 18(1):97–106
33. Shao J, Wu F, Ouyang C, Zhang X (2012) Sparse spectral hashing. Pattern Recog Lett 33(3):271–277
34. Smolensky P (1986) Information processing in dynamical systems: foundations of Harmony theory. In: Parallel distributed processing: explorations in the microstructure of cognition, MIT Press, vol. 1, pp 194–281
35. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
36. Wager S, Wang S, Liang PS (2013) Dropout training as adaptive regularization. In: Burges C, Bottou L, Welling M, Ghahramani Z, Weinberger K (eds) Advances in neural information processing systems vol. 26, pp 351–359
37. Wan L, Zeiler M, Zhang S, LeCun Y, Fergus R (2013) Regularization of neural networks using dropconnect. In: Proceedings of the 30th international conference on machine learning (ICML-13), pp 1058–1066
38. Wang J, Kumar S, Chang SF (2012) Semi-supervised hashing for large-scale search. Pattern Anal Mach Intell IEEE Trans 34(12):2393–2406
39. Weiss Y, Torralba A, Fergus R (2008) Spectral hashing. In: NIPS, vol. 9, p 6